图书情报工作
第63卷 第2期 2019年1月

中文电子病历的分词及实体识别研究

王若佳[1,2] 赵常煜[1] 王继民[1]
[1]北京大学信息管理系 北京 100871 [2]北京大学海洋研究院 北京 100871

摘要：［目的/意义］健康医疗大数据是我国重要的基础性战略资源，本研究对中文电子病历分词与实体识别的探讨与实证较好地完成了医疗数据的信息抽取任务，对今后医疗大数据在语义层面的应用发展具有重要意义。［方法/过程］本研究首先融合权威词表、官方标准、健康网站数据及其他医学补充词库构建了词语数量级达到10万的医学词表；然后对电子病历的字段进行分词，对比了jieba工具、导入词典后的jieba、无监督学习及AC自动机4种模型的分词效果；最后，以自动分词和人工标注结果为语料，实现基于条件随机场的电子病历实体识别研究，并比较不同实体类别以及不同文本特征下的实体识别效果，选出最优模板。［结果/结论］分词结果显示，AC自动机的效果最好，F值可达82%；实体识别结果表明，"检查"和"疾病"实体的识别效果最好，而"症状"的识别效果不太理想。

关键词：电子病历 中文分词 实体识别 健康医疗大数据 AC自动机 条件随机场
分类号：TP391.1
DOI：10.13266/j.issn.0252-3116.2019.02.004

1 引言

电子病历是医务人员在医疗活动过程中，使用信息系统生成的文字、符号、图表、图形、数字、影像等数字化信息，是居民个人在医疗机构就诊过程中产生和被记录的完整、详细的临床信息资源[1]。这些资源中包含大量潜在知识，对其进行挖掘，一方面可为医务人员提供临床决策的辅助和支持，另一方面可带来健康医疗模式的变化，提升健康医疗服务效率和质量。然而，我国健康医疗大数据研究仍处于早期阶段，电子病历中大部分信息以非结构化的文本形式保存，仅仅通过简单的分词无法识别与挖掘，需要针对医学文本中大量专业术语、众多新登录词以及结构化描述语言的特点，精进分词方法，并进一步识别医学实体。

鉴于此，本文首先从电子病历的中文分词入手，按照不同医学实体的类别，尽可能多并准地收集医学词汇，形成医学词典，探讨并比较基于词典方法、基于无监督学习方法以及多种方法的混合对中文病历的处理效果；然后基于条件随机场算法对电子病历中的疾病、症状、检查、药物和手术实体进行识别，完成病历的信息抽取任务。

2 文献综述

2.1 中文分词方法综述

分词是自然语言处理过程中最基础的语言处理模型之一。中文的语言构成比较复杂，难以进行准确词语识别。目前常见的中文分词方法主要包括两种：基于词典的语言匹配模型；基于统计和机器学习的计算模型。

2.1.1 基于词典的语言匹配模型 该方法是指词典中的每个词与被处理文档之间逐一匹配的过程。单词匹配时，常用的方法有最大匹配法，逆向最大匹配法，最小匹配法，最少切分法。这些方法的基本思路类似，主要策略是通过保存在编译的词典中的大量词条集来匹配被处理文档以找到可能的分词方式。

2.1.2 基于统计和机器学习的计算模型 该方法主要是通过相应模型来计算被分片段是否成为一个词组的概率。其中最常见的模型为隐马尔可夫模型和无监督分词模型。隐马尔可夫模型（Hidden marcov model，HMM）是发现新词或未登录词识别时用到一种方案。刘群等[2]通过Viterbi算法标注出全局最优的角色序列，然后在角色序列的基础上，利用HMM识别出未登录词，并计算出真实的可信度。李兆福[3]利用N-最

作者简介：王若佳（ORCID：0000-0003-1806-0688），博士研究生；赵常煜（ORCID：0000-0001-6780-1070），硕士；王继民（ORCID：0000-0002-3573-7788），教授，博士生导师，通讯作者，E-mail：wjm@pku.edu.cn。
收稿日期：2018-07-16 修回日期：2018-09-15 本文起止页码：34-42 本文责任编辑：杜杏叶

短路径算法,形成被测文本的有向图,降低了分词算法时间复杂度,为分词提供了新的思路。

目前中文分词工具和各方面的应用很多,如搜索引擎、语音识别技术、自动分类技术等。特别是2014年以来不少学者把分词技术应用于医学领域,对医学领域的文本进行分词并利用。从整体来看,在医学文本处理过程中经常出现许多医学专用词,如药品、疾病、身体器官、手术方法等。如果利用常用的分词工具进行处理,病历分词效果和识别率会大幅下降,有必要采取一定的措施对医学专用词进行识别。张立邦[4]利用半监督学习,包括有序聚类、期望极大算法(EM),对电子病历进行分词,并且从病历中进行了对象、药物等的命名实体识别。同年,张立邦等[5]又利用无监督的方式对电子病历进行了研究,在EM概率模型的基础上,通过信息熵构建良度,把未登录词识别问题转换为词语之间的最优化问题,利用动态规划算法进行分词结果的求解。此外,还有学者将词典和机器学习方法相结合,对电子病历进行分析。李国垒等[6]通过词典和统计相结合的分词算法,对出院记录进行了分词处理,并对临床术语和治疗方案进行潜在语义分词,对胃癌的分词方法进行了研究。

2.2 实体识别方法综述

由于本文的主要研究对象为临床电子病历,隶属于医学领域,因此主要对健康医疗领域中的实体识别方法做一总结。目前常用的实体识别方法主要包括以下两种:基于词典和规则的方法,基于机器学习的方法。

2.2.1 基于词典和规则的医学实体识别 基于规则的医学识别方法主要依靠术语词典及领域专家。早在1995年,哥伦比亚长老会医学中心(Columbia Presbyterian Medical Center, CPMC)的专家们设计出了MedLEE系统[7],该自然语言处理系统的主要任务是从病人报告中提取、结构化并编码临床信息,从而与临床信息系统相结合。MedLEE系统2012年还被M. Sevenster等学者用于从放射学报告中提取人体器官和临床发现之间的关系,并得到了82.32%-91.37%的准确度[8]。近20年间,MedLEE系统仍能取得良好效果的重要原因之一在于大型医学受控术语词典的支持。美国国立医学图书馆自1986年起开发的统一医学语言系统(Unified medical language system, UMLS)[9]以及CPMC在此基础上创建的医学实体词典(Medical entities dictionary, MED)[10]都为基于规则的医学实体识别奠定了基础。虽然这种方法较为原始,但颇有一些研究使用,因为有非常方便的开源工具支持,如MetaMap[9]、MedEx[10]、cTAKES[11],不过这些工具也仅针对英文文本,毕竟目前权威中文医学词典较少,规则设计也难以涵盖所有特例。

2.2.2 基于机器学习的医学实体识别 由于规则获取对专家的依赖性,越来越多的研究人员将重点放在了基于机器学习的医学实体识别上。Y. Li和S. L. Gorman[12]使用9 679份英文临床报告建立了隐马尔科夫模型,识别了报告中不同模块(如主诉、过敏情况、家族史、过往手术史等)的出现顺序;王鹏远和姬东鸿[13]分析了英文病历中的复合疾病问题,构建了多标签CRF模型。中文电子病历方面,叶枫等[14]采用条件随机场模型(Conditional random field, CRF),引入词性特征、构词模式特征、词边界特征和上下文特征,识别了250份中文病历中的疾病、临床症状和手术操作三类命名实体;J. Lei等人[15]比较了条件随机场、支持向量机(SVM)、最大熵(ME)和结构化支持向量机(SSVM)对400份中文入院记录和出院小结文本的实体识别效果进行分析,结果显示SSVM具有最高的F值(入院记录93.51%,出院小结90.01%);J. Liang等[16]提出了一种级联型中医药实体识别方法,即结合SVM和CRF算法的句子类别分类器,该分类器对中药的识别正确率高达94.2%,西药的评估F值也达到了91.7%。

综上所述,我们发现:①在电子病历分词方面,虽然分词技术在医学领域中的应用越来越重要,但是病历本身的非结构性及无统一格式等问题,分析效果难以提高,针对性比较强,整体缺乏通用的分析方法;②在电子病历实体识别方面,研究人员对病例的实体标注一般基于人工分词的方法完成,事实上不同人的分词标准很难统一,分词的细粒度对实体识别准确率影响较大。本研究将分词和实体识别视为两个关联任务,首先通过对比不同分词算法选出效果最好的,然后基于自动化分词结果进行下一步的人工标注工作。

3 研究设计与方法

3.1 研究设计

本研究的技术路线图见图1。整个研究流程可分为3个步骤:数据收集、实验研究、结果评估。首先融合权威词表、官方标准、健康网站数据及其他医学补充词库构建医学词表,同时从互联网上收集公开的中文电子病历数据,并对数据进行预处理。然后分别进行中文分词和医学实体识别的实验。在中文分词步骤中,对电子病历的字段进行分词,对比jieba工具、导入

词典后的 jieba、无监督学习及 AC 自动机 4 种模型的分词效果；在医学实体识别步骤中，以自动分词和人工标注结果为语料，实现基于条件随机场的电子病历实体识别研究。最后进行结果评估，比较分词结果以及不同文本特征下的实体识别效果，选出最优模板。

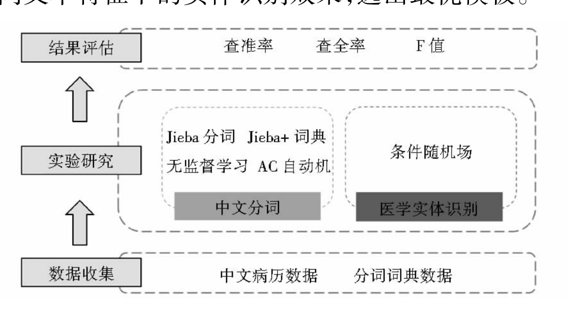


图 1　技术路线

3.2　数据来源

本研究使用的数据主要包括实验数据和词典数据两部分，前者用于分词效果评估和实体识别的训练及测试，后者用于构建分词词典。

3.2.1　实验数据　由于数据安全问题，本文的病历数据来自临床执业医师技能考试模拟题。该考试的第一站为病史采集与病例分析，其中，病例由患者性别、年龄、主诉、摘要（现病史、既往史及个人史）、检查（查体、辅助检查、实验室检查）、诊断、诊断依据、鉴别诊断、进一步检查和治疗原则共十部分组成。一条示例如表 1 所示，笔者于互联网中共采集类似病例 100 条（其中 90 条西医病例，10 条中医病例），用作分词和实体识别的实验数据。

表 1　病例分析示例数据

性别	男性	年龄	23 岁
主诉	因骑车进行中被汽车撞倒，右颞部着地半小时，到急诊就诊。		
摘要	患者摔倒后曾有约 5 分钟的昏迷，清醒后，自觉头痛，恶心。		
检查	BP 139 - 80mmHg，P80 次/分，一般情况可，神经系统检查未见阳性体征。头颅平片提示：右额颞线形骨折。遂将患者急诊留观。在随后 2 小时中，患者头疼逐渐加重，伴呕吐，烦燥不安，进而出现意识障碍。体检：T 38℃，BP 160/100mmHg，P60 次/分，R18 次/分，浅昏迷，左侧瞳孔 3mm，对光反射存在，右侧瞳孔 4mm，对光反应迟钝。左鼻唇沟浅，左侧 Babinski's Sign 阳性。		
诊断	右额颞急性硬膜外血肿		
诊断依据	有明确的外伤史；有典型的中间清醒期；头部受力点处有线形骨折；出现进行性颅内压增高并脑疝		
鉴别诊断	急性硬膜下血肿及颅内血肿：同有外伤史；血肿多出现于对冲部位；意识障碍持续加重；明确诊断靠 CT		
进一步检查	头颅 CT 平扫		
治疗原则	急诊行开颅血肿清除术		

注：病例来源为医学教育网（http://www.med66.com/zhiyeyishijineng/linchuangti/binglifenxi/）

3.2.2　词典数据　由于需要的是带有类别标注的分词词典，因此构建词典的第一步是定义类别。UMLS 对医学实体的定义最为权威且历史悠久，该系统定义的实体语义类型多达 133 种，包括解剖结构、生物功能、物理对象等 6 个大类[17]。不过对于医学病历来说，涉及不到如此多的实体类型。2010 i2b2/VA challenge 会议参考 UMLS 定义的语义类型，将电子病历命名实体分成了三类：医疗问题（Medical problem）、检查（Test）和治疗（Treatment）[18]。本文在此基础上将医疗问题拆分成疾病和症状，治疗拆分成药物和手术，最后得到 5 种医学实体类别——疾病、症状、检查、药物、手术。

构建词典的第二步是从以上 5 种类别出发，搜集与爬取每种类别的权威词表、官方网站、大众健康网站数据及其他补充词库，具体如下。

- 权威词表数据：目前中文医学权威词表包括基于美国国立医学图书馆《医学主题词表》中译本和《中国中医药学主题词表》的《中文医学主题词表》（CMeSH）、基于国际疾病分类 ICD - 10 扩展修订的国家标准《GB/T 14396 - 2016 疾病分类与代码》、国际疾病分类手术码 ICD - 9 - CM 中译版，以及中医药相关的《中医疾病国际标准编码》和《中医病证分类与代码》。
- 官方网站数据：由于缺少权威的中文药物名称词典及临床检验项目词典，本文下载并爬取了卫计委网站中的基本药物目录、医疗机构临床检验项目目录，以及国家食品药品监督局网站上药物列表中的所有药物名称。
- 健康网站数据：39 医学教育网（补网址）中的临床专业术语词典、检验参考值及临床意义、疾病临床科室分类以及手术库。
- 其他补充词库：百度百科医学词条，搜狗词库中较为可靠的医学词库。

构建词表的第三步是对已收集到的词典进行预处理。预处理的步骤包括：①为词语标注实体类别和来源词表；②删除分类代码、删除圆括号内的补充语句或

英文名称;③提取方括号内的补充同义词,添加为新词;④过滤出字符数大于20的词语,进行人工修改或删除。

最后,本文按以上步骤构建了一个包含136 253个词语的医学分词词典。其中,疾病类的词语数量为54 601个,症状类为3 316个,检查类为3 828个,药物类为57 390个,手术类为17 118个。

3.3 分词方法与原理

本研究使用常用分词工具jieba中文分词、AC自动机、无监督分词3种分词方式进行分词效果的对比。

jieba中文分词是最常见的中文分词工具之一,可提供中文分词、关键词抽取等多种功能[19]。jieba分词器以"最大匹配"作为匹配规则,完成基于词典的候选词挑选和最终结果的返回。假如利用词典搜索对于被测文本产生的候选词结果集合记为GEN(X),基于词典的模型可以表示为[20]:

$$Y = \mathrm{argmax}_{Y' \in GEN(X)} P(Y')$$

AC自动机(Aho and Corasick)是基于Trie树结构的多模式匹配算法[21],在信息检索,字符串匹配等多个领域有广泛的应用。AC自动机的步骤分为三步:①建立模式串的Trie树;②Trie树添加失败路径;③根据行程的自动机,搜索待处理的文本。整个搜索过程会出现当前关键词匹配和当前关键词不匹配两种情景:①当前关键词可匹配,表示从当前节点可以沿着树边有一条路径可以到达目标字符,此时只需沿该路径走向下一个节点继续匹配即可,目标字符串指针移向下一个关键词继续匹配;②当前字符不匹配时,去当前节点失败指针所指向的关键词继续匹配,匹配过程随着指针指向根节点结束。AC自动机需要重复上述过程中的一种,直到所有被测文档查到结尾为止。

本次研究中的无监督分词是利用凝固度的大小来判断是否成为一个词。一般凝固度大于一定值时,该片段看成一个词,接下来判断边界熵的问题。反过来,凝固度小于一定值时,该片段不能成为一个词。在被测文档中,把凝固度小于某一值的片段删除,剩下的片段则可以成为某一词。此过程,可以通过如下的数学方式表达。如果a、b是在被测文档中的相邻两个字,则可以统计(a、b)成对出现的次数F(a、b),以此可以估算该词的频率P(a、b)。然后,再对a、b分别统计出现次数F(a)、F(b)和出现频率P(a)、P(b)。如果满足以下公式则可以在语料中将两个字断开。

$$\frac{P(a,b)}{P(a)P(b)} < \alpha \text{ (α是给定的大于1的阈值)} \quad \text{式(1)}$$

通过上述方式,初步完成分词后,利用词频可以对候选词集进行筛选。

3.4 实体识别原理与方法

在自然语言处理领域中,条件随机场是一个序列标注算法,其结合了隐马尔科夫模型和最大熵模型的特点,不仅可以考虑词语本身和上下文特征,还可以加入词典等外部特征,具有较好的实体识别效果。简单来说,条件随机场是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型。不过一般用于标注时,都将其简化为线性链条件随机场,即输入变量和输出变量具有同样的结构,具体公式如下:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} q_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \quad \text{式(2)}$$

其中,Z(x)为规范化因子,计算公式为:

$$Z(x) = \sum_y \exp\left(\sum_{i,k} q_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \quad \text{式(3)}$$

在本研究中,x指病历文本序列变量,y为实体标注变量,P(y|x)是在给定x的条件下输出序列y的条件概率分布。前面提到,条件随机场可考虑多种文本特征,这是因为其有两类特征函数,第一类是状态特征函数$s_l(y_i, x, i)$,即实体类别只和当前病历文本有关;第二类是局部特征函数$t_k(y_{i-1}, y_i, x, i)$,其可考虑外部特征(如词性、大小写、上下文)对实体类别的影响。由此可见,病历实体识别的关键在于选取病历文本中的特征。本研究通过分析病历特点,采用了以下特征:

(1)词性特征:某位置的词语标注,既跟这个词的词性相关(如命名实体以名词居多),也可能跟上下文词语的词性相关(如疾病实体经常出现在"患""诊断"等动词前后)。本研究借助jieba分词工具进行词性标注工作。

(2)上下文特征:病历文本本身存在的内在规律和特点,需要选择"上下文窗口"构成长度,比如对于下句"因 骑车 进行中 被 汽车 撞倒 右颞部 着地 半小时 到 急诊 就诊",假设此刻为i,所在位置的词语为"右颞部",窗口长度为5,则算法会自动提取i－2,i－1,i,i＋1,i＋2这5个时刻的词语,构成一个长度为5的窗口,如图2所示:

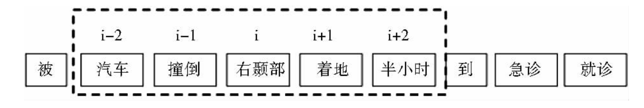


图2 上下文特征与滑动窗口长度

（3）所属病历模块：在我们的病历样本中，每个病历都可分为性别、年龄、主诉、摘要、检查、诊断等 10 个模块，分析发现不同实体类别在不同模块中出现的频次有很大不同，例如诊断及鉴别诊断模块中出现的实体多为疾病，治疗模块中则常见药物和手术实体。因此对于每个词语，为其标注所属模块，模块名称与标识符号的对应关系如表 2 所示：

表 2　模块名称与标识符号对应关系

模块名称	标识符号	模块名称	标识符号
性别	XB	诊断	ZD
年龄	NL	诊断依据	YJ
主诉	ZS	鉴别诊断	JB
摘要	ZY	进一步检查	JJ
检查	JC	治疗原则	ZL

（4）模块中的位置：词语所在病历模块中的相对位置可能会反映其类别。例如，在主诉模块中，大部分句子的构成特征为“症状 + 程度”（如腹痛 2 小时，呕吐 3 次）。计算该词语在模块中出现的位置顺序数和模块中共有多少个词语数，二者相比，则可以用一个取值区间为[0，1]的小数来标识每个词语在模块中的位置。

本研究使用 CRF + + 工具实现基于条件随机场算法的实体标注。该工具需要用户导入训练文件和测试文件，并依据所选特征构建模板文件。一个训练文件的示例见图 3。第一列为病历文本，第二列为词语词性，第三列为所在病历模块标识，第四列为模块中的相对位置，最后为实体标注类别。我们采用 BMES 标注法（B 是词首，M 是词中，E 是词尾，S 为无关词）。

因	p	ZS	0.077	S
骑车	n	ZS	0.154	S
进行	v	ZS	0.231	S
中	f	ZS	0.308	S
被	p	ZS	0.385	S
汽车	n	ZS	0.462	S
撞倒	v	ZS	0.538	S
右颞部	n	ZS	0.615	B_Symptom
着地	x	ZS	0.692	E_Symptom
半小时	n	ZS	0.769	S
到	v	ZS	0.846	S
急诊	v	ZS	0.923	S
就诊	v	ZS	1.000	S

图 3　一个训练文件的示例样本

CRF + + 的模板文件可描述病历上下文信息，并选择其他特征。本研究主要采用一元特征（Unigram）模板形式，定义的模板特征见表 3。其中，类似于“U00”的符号为特征编号，书写特征时采用% x[row，col]的形式，row 表示与当前位置的相对行数，col 表示与当前位置的相对列数。

表 3　电子病历实体识别时的特征模板定义

特征名称	特征序号	模板书写	模板含义
上下文特征	U00 - 04	% x[-2,0]、% x[-1,0]、% x[0,0]、% x[1,0]、% x[2,0]	窗口长度为 1
	U05 - 06	% x[-1,0]/% x[0,0]、% x[0,0]/% x[1,0]	窗口长度为 2
	U07	% x[-1,0]/% x[0,0]/% x[1,0]	窗口长度为 3
	U08	% x[-2,0]/% x[-1,0]/% x[0,0] /% x[1,0]/% x[2,0]	窗口长度为 5
词性特征	U09	% x[0, 1]	此位置的词的词性
	U10	% x[-1, 1]	上一位置词的词性
	U11	% x[1, 1]	下一位置词的词性
	U12	% x[-2, 1]	上两位置词的词性
	U13	% x[2, 1]	下两位置词的词性
模块特征	U14	% x[0, 2]	此位置的词所属模块
	U15	% x[-1, 2]	上一位置词所属模块
	U16	% x[1, 2]	下一位置词所属模块
位置特征	U17	% x[0, 3]	此位置的词在所属模块中的相对位置

3.4　结果评估方法

对结果评估方法通常采用召回率（recall）、准确率（precision）、F 值等测评指标。在中文分词和实体识别过程中都会采用以上 3 种指标，具体如下：

（1）中文分词评估指标。

$$\text{分词召回率(recall)} = \frac{\text{算法正确切分词语总数}}{\text{人工切分词语的总数}} \times 100\%$$

$$\text{分词准确率(precision)} = \frac{\text{算法正确切分词语总数}}{\text{算法切分的词语总数}} \times 100\%$$

$$\text{分词 F 值} = \frac{2 \times \text{分词召回率} \times \text{分词准确率}}{\text{分词召回率} + \text{分词准确率}} \times 100\%$$

（2）实体识别评估指标。

$$\text{实体识别召回率(recall)} = \left(\sum_{i=1}^{n} \frac{\text{算法正确识别的 i 类词性总数}}{\text{人工识别的 i 类词性总数}}\right) / (n) \times 100\%$$

实体识别准确率(precision)＝

$$\left(\sum_{i=1}^{n}\frac{\text{算法正确识别的 i 类词性总数}}{\text{算法切分的 i 类词性总数}}\right)/(n)\times 100\%$$

实体识别 F 值＝

$$\frac{2\times\text{实体识别召回率}\times\text{实体识别准确率}}{\text{实体识别召回率}+\text{实体识别准确率}}\times 100\%$$

召回率和准确率越高，表明分词和实体识别算法的效果越理想。但是召回率和准确率，如同查全率和查准率，相互互斥，通常一项较高，则另一项较低[22]。为了综合考虑召回率和准确率，将把两项指标综合计算为 F 值，对算法进行测评。F 值越高表明该算法的性能越好。本次实验以人工分词和人工标识为标准，人工分词和标识过程采取多人相互验证的方式。

4 研究结果

4.1 分词结果

本研究对比了 jieba 分词、jieba＋用户词典、无监督学习及 AC 自动机四种策略对病历文本的分词效果。分词原则如表 4 所示：

表 4 电子病历分词原则

分词原则	分词方法			
方法名称	jieba 分词	jieba＋用户词典	无监督学习	AC 自动机
是否用词典	否	是	否	是
核心算法	$Y = \text{argmax}_{Y' \in GEN(X)} P(Y')$		$\frac{P(a,b)}{P(a)P(b)} < \alpha$	Trie 结构
是否涉及第三方库	是	是	否	是

按照上述 4 种方案对电子病历进行分词，结果见表 5。其中，AC 自动机的 F 值最高，可以达到 82%；其次为加入医学词典的 jieba 分词，F 值为 74%；第三为基于信息熵方法的无监督学习模型，F 值为 69%；最后为没有词典的 jieba 分词器，F 值为 66%。

表 5 电子病历分词结果原则

测评结果	召回率	准确率	F 值
jieba 分词	72%	60%	66%
jieba＋用户词典	81%	68%	74%
无监督学习	78%	62%	69%
AC 自动机	89%	76%	82%

出现以上结果的原因，可以归结为以下几个方面：

(1)电子病历数量的影响。本研究的样本数据为 100 条电子病历，整体样本量偏低，导致无监督学习和不涉及到词典的算法的召回率和准确率都偏低。在无监督学习过程中，样本量较少，难以进行多项边界熵交叉分析；而基于词典的算法不依赖于样本数，而依赖于词典的准确性。由于收集到的词语数量和范围比较广，因此基于词典的分词效果更为理想。

(2)数量单位的敏感度影响。电子病历中，尤其是实验室检查部分涉及很多数量单位，如人体体温的摄氏度，用药量的 U/L、体内细胞含量的 μmol/L 等。四种分词算法对数量单位的敏感度不同，导致整体效果的不同。涉及到词典的算法对数量单位的敏感度比较强，单位的区分度比较好，而基于无监督学习的分词，单位敏感度比较低。如“U/L”和“μmol/L”，无监督学习过程中，总会把“/L”部分和前半部分分开，其结果为“U”和“/L”，“μmol”和“/L”。

(3)医学词语的专业性较强。由于不同患者的病史和治疗方案不同，病历中涉及到的医学词语多种多样。一些词语在病历中的出现次数较少，词频较低，一定程度上导致无监督学习的效果比较差。总而言之，由于医学病历的专业性比较强，光靠无监督学习难以进行准确的分词。结合词典和统计学习的方式，可以大大提高分词的效果。

4.2 实体识别结果

本研究在构建训练样本和测试样本时，采取层次抽样的方法，将中、西医病历分开，每种类别病历中各抽取 70% 用作训练，余下 30% 为测试样本。其中，训练样本中共有 14 664 个词，测试样本有 6 028 个词。整体来看，模型效果如表 6 所示：

表 6 病历实体识别效果评估

评估指标	数值
算法识别的正确实体数	1 846
算法识别的所有实体数	2 119
实际实体数	2 460
准确率	87.12%
召回率	75.04%
F 值	80.63%

4.2.1 不同类别对实体识别结果的影响 不同类别的实体对识别结果也会有一定影响。从 F 值看(见图 4)，“检查”和“疾病”的识别效果最好，其次是“药物”与“手术”，“症状”的识别效果不太理想。

对比准确率和召回率(见图 5)，所有实体类别的准确率都明显高于召回率，其中以“药物”实体最为明显，准确率高达 94%，召回率却只有 63%；其次是“症状”，即便准确率为 88%，由于召回率最低(62%)，导致 F 值不甚理想；值得一提的是“检查”实体的识别结果，准确率与召回率相当，都在 85% 左右。

4.2.2 不同特征对实体识别结果的影响 特征选择对 CRF 的训练效果具有很大影响。为了逐步分析不

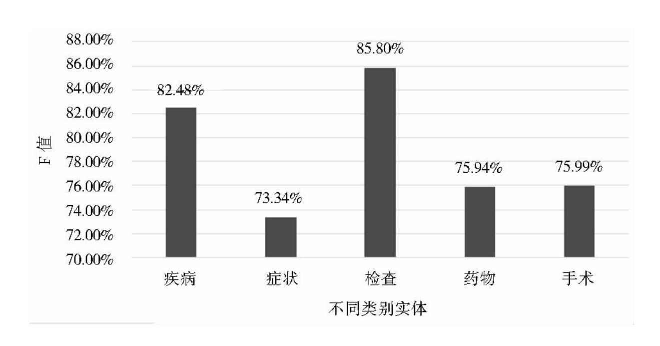


图 4　不同类别的实体识别 F 值结果

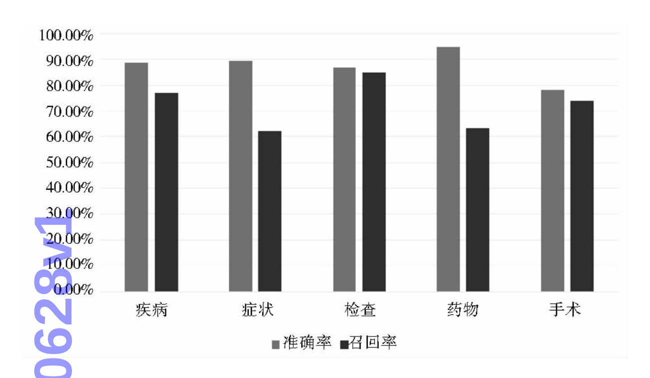


图 5　不同类别的实体识别准确率与召回率比较

同特征和特征组合对于 CRF 自动识别的影响，寻求相对最佳的模板设置，本研究分别对比了仅选取一个特征，任选两个特征和选取所有特征的实体识别效果。

首先，由于位置特征和模块特征息息相关，为探讨位置特征是否对实体识别有所帮助，需要对比仅有模块特征以及模块 + 位置特征的识别效果。结果显示（见表 7），位置特征不仅没有增进模型的识别效果，反而大大拉低准确率与召回率，因此后续将不予考虑此特征。

表 7　模块特征与“模块 + 位置”特征的识别效果对比

	准确率	召回率	F 值
只有模块特征	87.85%	67.28%	76.20%
模块 + 位置特征	49.13%	33.33%	39.72%

仅选取一个特征的对比结果见表 8，可以看到上下文特征和模块特征的 F 值更高，其中上下文特征的识别准确率达到 90%。

表 8　仅选取一个特征的识别效果对比

	准确率	召回率	F 值
上下文特征	90.83%	61.18%	73.11%
词性特征	47.15%	30.65%	37.15%
模块特征	87.85%	67.28%	76.20%

任选两个特征的对比结果见表 9，词性 + 模块特征的 F 值最高，甚至优于考虑到所有特征的模型效果。可见一味追求融合更多特征并不是关键。

表 9　任选两个特征的识别效果对比

	准确率	召回率	F 值
上下文 + 词性	86.48%	69.67%	77.17%
上下文 + 模块	89.04%	71.34%	79.21%
词性 + 模块	87.78%	78.54%	82.90%

5　讨论与结论

5.1　研究结果

为响应健康医疗大数据应用发展的需求，本研究对医疗活动中最常见的电子病历数据进行了分词和实体识别，得到的研究结果如下：

（1）AC 自动机的效果最好，其 F 值可达 82%。原因在于该算法融合了词典和统计方法，一方面本身基于大量的医学词典，对特殊的医学词条比较敏感，另一方面结合了统计方法，可以对未登录词进行挖掘和发现。此外，电子病历数量的影响、数量单位的敏感度影响，以及医学词语的强专业性，导致无监督学习的方式和开源分词工具的分词效果较差。总之，针对专业性比较强的领域，结合词典的方式可以一定程度上提高分词的效果。

（2）基于条件随机场的病历实体识别效果较好，F 值最高为 82.9%。对于不同类别的实体来说，“检查”和“疾病”的识别效果最好，因为二者具有很强的格式化特征，病历中的常见检查项目高度相似，而疾病实体多出自病历的“诊断”和“鉴别诊断”两个模块；“症状”的识别效果不太理想，一方面在于描述症状的语句较为口语化，另一方面与其出现在多个病历模块中有关；病历文本的位置特征并没有增进模型的识别效果，并不一定是因为该特征对实体识别的贡献不大，有可能是本研究定义的数字表达方式造成的机器误解；特征选取不在于数量多少，能更准确地表达文本语义才是研究的关键。

5.2　研究不足

在研究的过程中，我们遇到了以下问题：

（1）词典的筛选、合并与构建。目前词典构建的来源数据除了官方、权威数据外，还有来自网络的医学词库，虽然网络词库可大大扩充词语数量级，但其正确性不敢保证。权威医学词表虽然很正式，但预处理非常繁琐，且很有可能出现数据处理不当的问题。例如 ICD－10 中，大部分方括号中的内容为可替换的同义词，因此我们采取“提取方括号内词语，添加为新词”的处理方法，这种做法对于“利斯特菌病[李司忒氏菌病]”这个词语来说，可直接提取出[李司忒氏菌病]，

但对于“中型[典型]霍乱”，所提取出的[典型]一词并不是一个疾病；再比如，ICD－9－CM 中，有很多辅助编码词，“其他淋巴结根治性切除术排除：合并根治性乳房切除”中的“排除”以及“胸部食管造口术另编码：任何部分切除术”中的“另编码”都不是手术中的术语，处理起来问题很多。

（2）人工标注的规则制定。由于目前缺乏中文的电子病历人工标注语料库，因此需要我们的研究小组自行标注。然而，医学领域非常特殊，非专业人员理解起来尚有困难，虽然小组中有一位医学本科背景的成员，但和临床医生的经验相去甚远。人工标注的规则制定，一方面需要请教医生，比如“骨折”“休克”“药物过敏”等词究竟为疾病还是症状；另一方面需详细讨论分词的细粒度问题，如“自觉头痛”应分为两个词“自觉 头痛”还是一个词，“心、肺、腹未见异常”应分为“心肺腹 未见 异常”还是“心 肺 腹 未 见 异常”等。

（3）医学实体的修饰问题。在标注的过程中，我们发现即便机器学习对了，但在语义上也有可能千差万别。例如“3 周前咽部不适”中的时间成分“3 周”，说明了该症状的时间与程度；“肺结核（浸润型？慢性纤维空洞型？）”中的待查成分“？”，说明并不确定该患者患有哪种结核病；最重要的是类似于“巩膜无黄染”中的否定成分“无”，“巩膜黄染”与“无黄染”对医生的诊断影响很大，若仅识别并提取出“黄染”二字，对患者的健康状况度量有重要影响，对今后进一步的数据挖掘甚至会产生反面作用。

5.3 未来研究方向

之后的研究将先解决如前所述的不足与问题，完善词典，咨询、调查与构建人工标注规则，定义并提取常见医学实体修饰。此外，还有两个方面的发展方向。

（1）实体标注的半自动化实现。在本研究中，对实体的类别标注主要通过人工完成，事实上耗费了较多的时间与精力。今后，可探讨如何将医学词典应用于标注实体的过程中，或者构建多分类器实现协同训练，从而尽可能地减少人工标注的工作量。

（2）实体关系的定义与抽取。电子病历中的实体不是孤立存在的，相互之间存在着一定的关系，这种关系正是医疗知识的主要体现。因此，在识别命名实体的基础上，还需要定义不同实体之间的关系，从而形成医疗领域的知识图谱。

参考文献：

[1] 国家卫生健康委员会. 电子病历应用管理规范（试行）[EB/OL]. [2018－02－20]. http://www.nhfpc.gov.cn/yzygj/s3593/201702/22bb2525318f496f846e8566754876a1.shtml.

[2] 刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展，2004，41(8):1421－1429.

[3] 李兆福. 基于 K 最短路径的中文分词算法研究与实现[D]. 哈尔滨:哈尔滨工程大学，2009.

[4] 张立邦. 基于半监督学习的中文电子病历分词和名实体挖掘[D]. 哈尔滨:哈尔滨工业大学，2014.

[5] 张立邦，关毅，杨锦峰. 基于无监督学习的中文电子病历分词[J]. 智能计算机与应用，2014(2):68－71.

[6] 李国垒，陈先来，夏冬，等. 面向临床决策的电子病历文本潜在语义分析[J]. 现代图书情报技术，2016，32(3):50－57.

[7] FRIEDMAN C，HRIPCSAK G，DUMOUCHEL W，et al. Natural language processing in an operational clinical information system[J]. Natural language engineering，1995，1(1):83－108.

[8] SEVENSTER M，VAN O R，QIAN Y. Automatically correlating clinical findings and body locations in radiology reports using MedLEE[J]. Journal of digital imaging，2012，25(2):240－249.

[9] MetaMap. A Tool For Recognizing UMLS Concepts in Text [EB/OL]. [2018－08－18]. https://mmtx.nlm.nih.gov/.

[10] XU H，STENNER S P，DOAN S，et al. MedEx：a medication information extraction system for clinical narratives[J]. Journal of the American medical informatics association，2010，17(1):19－24.

[11] SAVOVA G K，MASANZ J J，OGREN P V，et al. Mayo clinical text analysis and knowledge extraction system (cTAKES)：architecture，component evaluation and applications[J]. Journal of the American medical informatics association jamia，2010，17(5):507－513.

[12] LI Y，GORMAN S L. Section classification in clinical notes using supervised hidden markov model[C]// Arlington，VA，USA:Proceedings of the 1st ACM International Health Informatics Symposium. ACM，2010:744－750.

[13] 王鹏远，姬东鸿. 基于多标签 CRF 的疾病名称抽取[J]. 计算机应用研究，2017，34(1):118－122.

[14] 叶枫，陈莺莺，周根贵，等. 电子病历中命名实体的智能识别[J]. 中国生物医学工程学报，2011，30(2):256－262.

[15] LEI J，TANG B，LU X，et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American medical informatics association，2014，21(5):808－814.

[16] LIANG J，XIAN X，HE X，et al. A novel approach towards medical entity recognition in Chinese clinical text[J]. Journal of healthcare engineering,2017(2):1－16.

[17] UMLS. Current semantic types[EB/OL]. [2018－02－20]. https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html.

[18] UZUNER Ö，SOUTH B R，SHEN S Y，et al. 2010 i2b2/VA challenge on concepts，assertions，and relations in clinical text[J]. Journal of the American medical informatics association，2011，18

(5):552-556.
[19] 结巴中文分词[EB/OL]. [2018-02-20]. https://github.com.
[20] 沈翔翔, 李小勇. 使用无监督学习改进中文分词[J]. 小型微型计算机系统, 2017, 38(4):744-748.
[21] 孔东林, 罗向阳, 邓崎皓, 等. 基于AC自动机匹配算法的入侵检测系统研究[J]. 微电子学与计算机, 2005, 22(3):89-92.
[22] 李原. 中文文本分类中分词和特征选择方法研究[D]. 长春:吉林大学, 2011.

作者贡献说明:
王若佳:确定研究框架,收集病例,完成实体识别部分的综述、实验及撰写;
赵常煜:爬取并整理词表,完成分词部分综述、实验及撰写;
王继民:确定论文选题,论文指导与修改。

Healthcare Data Mining: Word Segmentation and Named Entity Recognition in Chinese Electronic Medical Record

Wang Ruojia[1,2] Cho Sang Wouk[1] Wang Jimin[1]
[1] Department of information management, Peking University, Beijing 100871
[2] Institute of Ocean Research, Peking University, Beijing 100871

Abstract: [**Purpose/significance**] Healthcare big data is an important basic strategic resource in China. Word segmentation and entity recognition of Chinese electronic medical record (EMR) is helpful in extracting important information from a large number of unstructured text. [**Method/process**] In this study, a Chinese medical thesaurus is firstly built in terms of authoritative medical subject headings, official standards and health website data; then, the effect of four segmentation methods is compared based on the corpus of artificial segmentation and manual annotation; finally, CRF model is used to identify 5 entities, including disease, symptom, test, drug and treatment. [**Result/conclusion**] Results show that (i) AC automaton model has the best F-measure in EMR word segmentation, which is 82%; (ii) compared with Western medical record, it's difficult to identify medical entities in the record of traditional Chinese medicine. Besides, "Test" and "Disease" entities have better F-measure, while the F-measure of "Symptom" entity is not that ideal.

Keywords: healthcare data mining electronic medical record Chinese word segmentation named entity recognition AC automaton conditional random field

下 期 要 目

□ 数据驱动下数字图书馆知识发现服务优化与应用研究 (毕强)
□ 云计算环境下学术信息资源共享全面安全保障机制 (石宇 胡昌平)
□ 应用生物反馈技术的焦虑症图书馆阅读疗法研究 (杨桦 卢章平 袁润)
□ 全面质量管理视角下的阅读推广研究 (张欢 谭英 夏圆)
□ 跨学科团队的信息交流规律研究:以威斯康辛麦迪逊分校为例 (李晶 章彰 张帅)
□ 欧洲科研开放获取基础设施项目OpenAIRE的建设与启示 (孙茜)